A PROTEOMIC APPROACH FOR IDENTIFICATION OF BACTERIA USING TANDEM MASS SPECTROMETRY COMBINED WITH A TRANSLATOME DATABASE AND STATISTICAL SCORING

Jacek P. Dworzanski
Geo-Centers, Inc., Aberdeen Proving Ground, MD 21010-0068
A. Peter Snyder
U.S. Army Edgewood Chemical Biological Center, Aberdeen Proving Ground, MD 21010-5424
Haiyan Zhang and David Wishart
Faculty of Pharmacy and Pharmaceutical Sciences, University of Alberta, Edmonton, Alberta T6G2N8,CA
Rui Chen and Liang Li
Department of Chemistry, University of Alberta, Edmonton, Alberta T6G 2G2, CA

ABSTRACT

We describe a novel method for fast identification of bacteria based on amino acid sequencing of a random set of peptides derived from bacterial cellular proteins through searches of their product ion mass spectra against a database translated from fully sequenced bacterial genomes. An in-house developed algorithm for filtering of search results have been tested with *Bacillus subtilis* and *Escherichia coli* microorganisms. This approach allowed for the selection of peptide assignments to microbial genomes with desired specificities, sensitivities and error rates, allowing for unambiguous identification of the test bacteria among 87 database microorganisms.

INTRODUCTION

Fast detection and identification of pathogenic agents of biological origin including viruses, bacteria and toxins play a crucial role in a proper response to unintentional or terrorist caused outbreaks of infectious diseases and the use of biological warfare agents on the battlefield. Recently, genomes of all bacteria listed as priority bacterial pathogens for biodefense purposes[1] have been sequenced and this achievement opens new possibilities for fast and reliable identification of these bacteria on a molecular level by retrieving their genomic information. We propose a genome-wide assay to probe genomic DNA sequences of microorganisms via amino acid sequencing of a random set of peptides derived from expressed proteins by using mass spectrometry (MS) and proteomics.

Currently there is an urgent need to develop fast and reliable methods to retrieve parts of genomic information that are thought to be representative of the whole genome. The growing number of completely sequenced bacterial genomes,[2] and as a result the availability of genomic databases for almost one hundred bacteria, provides the sequence information of every potentially expressed protein encoded by these organisms.

Several groups have reported the application of MS to obtain partial protein sequence information for the purpose of microorganism identification. Yates and Eng[3] claimed a method for identifying an organism of interest by determining whether tandem mass spectra of peptides obtained from its proteins indicate a homology to a portion of any proteins specified by amino acid sequences in a library of known

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
01 OCT 2005	**N/A**	**-**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
A Proteomic Approach For Identification Of Bacteria Using Tandem Mass Spectrometry Combined With A Translatome Database And Statistical Scoring	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Geo-Centers, Inc., Aberdeen Proving Ground, MD 21010-0068**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES
See also ADM001851, Proceedings of the 2003 Joint Service Scientific Conference on Chemical & Biological Defense Research, 17-20 November 2003. , The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**UU**	**7**	

organisms. Harris and Reilly[4] demonstrated that searching matrix-assisted laser desorption/ionization post-source decay mass spectra of peptides derived from on-probe digested *B. subtilis* against only the *B. subtilis* SwissProt database produces matches to particular proteins. Yao et al.[5] found that searching a SwissProt database using tandem mass spectra of a few peptides from limited proteolytic digestion of a virus provided high scoring matches to proteins coded by the subject virus and on this basis claimed its identification.

We have developed an electrospray ionization (ESI) tandem MS (MS/MS) method for the identification of bacteria represented in a database. In this method, which is experimentally similar to the shotgun proteomics work by Yates and co-workers,[3,6,7] proteins released during cell lysis are first digested, followed by one-dimensional (1D) reverse phase (RP)-liquid chromatography (LC) separation and ESI-MS/MS analysis of the resulting peptides. However, instead of searching the tandem mass spectra data against the genome database of microorganisms, we have constructed a translated virtual bacterial proteome ('translatome') database from 87 complete bacterial genomes by finding protein coding genes and translating codons into amino acid sequences of all potentially expressed proteins. In addition, we have developed a statistical scoring algorithm to rank the matches of the experimental data generated by the SEQUEST searching engine to identify a bacterium with high confidence.

EXPERIMENTAL SECTION

Processing of Bactrial Cells. *E. coli K-12* (ATCC 47076, Ec) and *B. subtilis* (ATCC 23857, Bs) strains were obtained from the American Type Culture Collection. Bacterial cells were incubated under ATCC recommended conditions, harvested, washed with distilled water, lyophilized, and stored at -25°C until analysis. Proteins were extracted from bacterial cells after lysis with sonication (Branson probe sonicator; Branson Ultrasonics Corp., Danbury, CT) in 100 mM ammonium bicarbonate buffer (pH 8.5) for 2 minutes. The resulting suspensions were centrifuged at 11,750 *g*. The supernatants were then filtered using Microcon-3 filters (Millipore, Mississauga, ON) with a 3000 Da molecular mass cut-off. The cell extract was denatured with urea, reduced with dithiothreitol, and digested by trypsin at a ratio of 1:50 (w/w). A Zip Tip C18 was used to desalt the resulting peptide mixture (Millipore, Mississauga, ON) before HPLC analysis. All reagents were from Sigma (St. Louis, MO). Distilled water was from a Milli-Q UV plus ultra-pure system (Millipore, Mississauga, ON).

LC and Acquisition of Mass and Tandem Mass Spectra. 1D LC-MS/MS was conducted on an LCQ DECA Surveyor LC-MS system (ThermoFinnigan, San Jose, CA). Chromatographic separation was performed on a Vydac C18 column (300 Å, 5 μm, 150 μm i.d. × 150 mm) with a flow rate of 1 μL/min. The mobile phase consisted of water and MeCN, and both contained 0.5% (v/v) acetic acid.

For two-dimensional (2D) LC-MS/MS, the peptide mixtures were first separated on a Vydac sulfonic acid cation-exchange column (900 Å, 8 μm, 300 μm i.d. × 150 mm) using a step-gradient with increasing NaCl concentration from 0 to 500 mM. Solvent delivery was performed on an Agilent (Palo Alto, CA) HP 1100 HPLC system. Two alternating Vydac C18 columns in an automated fashion separated the effluent from the first dimension Vydac column.

The LCQ DECA ion trap was set to acquire a full mass spectral scan between m/z 400 and 1400 followed by full tandem mass spectral scans between m/z 400 and 2000 of the three most intense ions from the survey scan. Relative collision energy for collision-induced dissociation (CID) was set to 35% with a 30 ms activation time. Dynamic exclusion was enabled with a repeat duration of 0.5 min, repeat count of 2, and a 3 min exclusion duration window.

Database Construction. Complete genome sequences of 87 bacteria which were available during our studies were downloaded from the National Center for Biotechnology Information (NCBI) site[8] and automatically processed. A computational Gene Locator and Interpolated Markov Modeler (GLIMMER 2.0), made available by The Institute for Genomic Research (TIGR, Rockville, MD), was used to recognize protein coding regions and to identify open reading frames (ORFs).[9] In-house written software was applied for automatic translation of these codons into amino acid sequences of all putative proteins and to assemble a protein database. The database was additionally processed to create a 'virtual peptide translatome' of all potential tryptic peptides. These translated sequences were searched by tandem a mass

spectra data mining software SEQUEST. The total of 296,942 protein-encoding putative genes recognized by GLIMMER in 87 bacterial genomes were used for the database construction.

Data Processing. The results of searches of product ion mass spectra against the in-house database using the SEQUEST algorithm[10] produces assignments of matching peptides to translated bacterial proteomes (translatomes). The validity of fit for each assignment was estimated using discriminant analysis applied to the training dataset of a combined 3,019 sequencing attempts. The tandem mass spectra were obtained from 17 2D-HPLC-peptide separations coupled with nano-ESI-MS/MS analyses of an *E. coli* digest. Spectra corresponding to tryptic peptide sequences of putative proteins in the *E. coli* K-12 database were considered as correct assignments while all the remaining matches as incorrect. Discriminant analysis and modeling of discriminant score distributions among correct and incorrect peptide assignments were performed using Statistica software (release 6, StatSoft, Inc., Tulsa, OK) and included five SEQUEST output parameters: Xcorr, ΔCn, Sp, RSp and ΔMpep as variables. ΔMpep refers to the absolute value of the mass difference between a peptide characterized by a molecular ion with a postulated charge state in the mass spectrum and the theoretical mass of the assigned peptide.

RESULTS AND DISCUSSION

Figure 1 shows the schematic representation of the proteomic approach for identification of bacteria based on MS/MS analysis of a whole cell protein digest, database search and statistical analysis of the matching scores. To obtain product ion mass spectra of tryptic peptides a typical shotgun proteomics procedure involving extraction of cellular proteins, digestion, and HPLC-MS/MS analysis of peptides was used. These product ion mass spectra are searched against a database composed of genome-translated proteomes (i.e., translatomes) of microorganisms and the searching results are analyzed using in-house developed software for statistical scoring of peptide assignments to translatomes and bacteria identification.

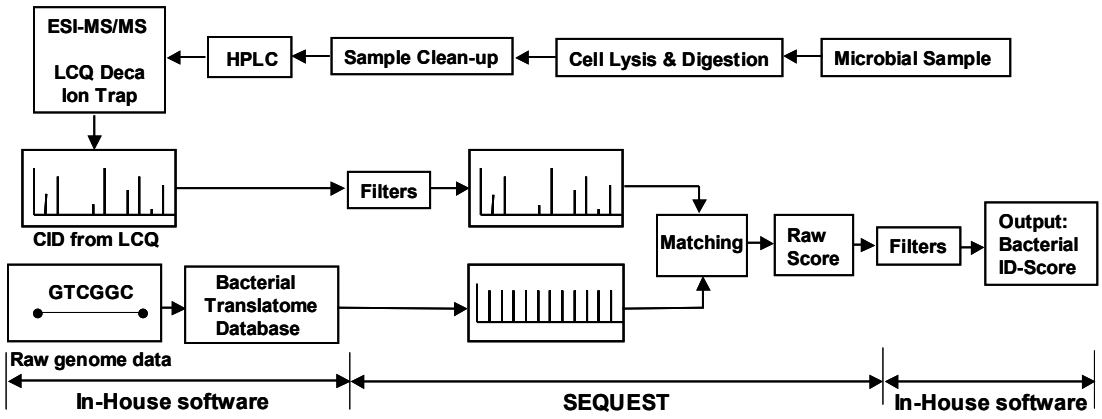

Figure 1. Schematic representation of the experimental set-up and data processing used for the identification of bacteria.

Although the experimental scheme used in this report was developed previously for identification of specific proteins, the present method does not relies on the identification of particular proteins but emphasizes amino acid sequences unique for a given bacterium. To indicate the presence of such amino acid strings in the protein database we suggest the term 'translatome' as appropriate terminology, reflecting the use of sequences translated from of all nascent hypothetical proteins coded in analyzed genomes.

Modeling of Database Search Scores. The approach used in this work to distinguish correct and incorrect peptide matches was based on modeling of the SEQUEST computed scores using a multivariate discriminant function (DF) analysis.[11] DF analysis transforms SEQUEST scores into DF scores and maximizes the ratio of between-class variance to within-class variance by computing appropriate weights

associated with each variable. The results of discriminant analysis used to evaluate 3019 sequencing attempts from product ion mass spectra obtained by 2D HPLC-MS/MS analysis of an *E. coli* tryptic digest are displayed in Figure 2 A. Two categories of assignments were used: (1) correctly identified peptides that represent an informational signal and (2) incorrectly identified peptides that can be treated as noise derived from random matches. Therefore, for each sequencing attempt a discriminant function score, calculated using the equation DF = 0.595 Xcorr + 6.620 ΔCn - 0.0001 Sp – 0.237 ln(RSp) –

– 0.134 ΔM_{pep} – 0.77, replaced a set of parameters generated by SEQUEST.

Distributions of these DF scores are shown in Figure 2A as approximations modeled for incorrect and correct peptide assignments by the log-normal and Gaussian distribution functions, respectively. A user selected decision point (vertical line in Figure 2A) represents a decision criterion associated with the tradeoff between sensitivity and specificity of a test.

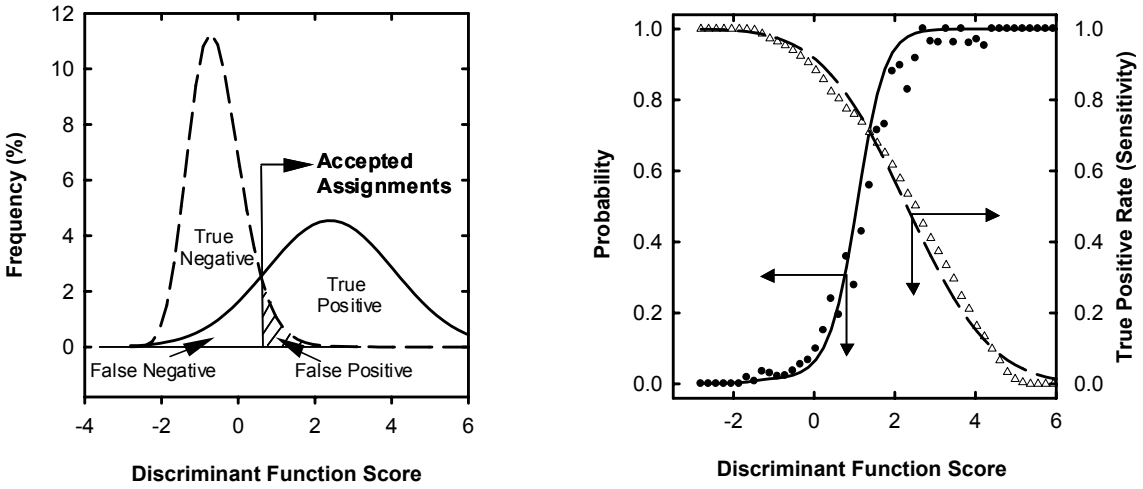

Figure 2. (A) Normalized distributions of DF scores for a training dataset of positively (—, correct assignments) and negatively assigned peptides (**---,** incorrect assignments). A vertical solid line indicates an example of a decision criterion that divides assignments into accepted (true and false positive) and rejected matches (true and false negative).

(B) Probability and sensitivity of correct peptide identification: the observed probabilities (•) represent the fraction of correctly matched peptides for each bin of DF scores, while the expected probabilities (—) were calculated using modeled distributions of correct and incorrect matches; sensitivities were calculated from the relationship (1 – cumulative fraction of correct peptide assignments) and are displayed as the observed true positive rates (Δ) and rates computed from modeled distributions (---).

Distributions experimentally observed and approximated using modeling functions were used to calculate the observed and expected probabilities that a peptide is correctly identified. The results are presented in Figure 2B, and the probability curve illustrates how DF scores can be translated into the probability of correct peptide identification. This allows a probability based decision with respect to peptides accepted for further evaluation. Note that the higher the probability that accepted peptides are correctly identified, the lower the fraction of correctly identified peptides are actually included for identification purposes (sensitivity). There are a number of reasons that cause poor matches between recorded product ion mass spectra and the respective theoretical spectra of peptides that are both of biological and methodological origin. The biological reasons include mainly mutations specific for a given strain while the methodological errors include those associated with the preparation of the database, the sample preparation process, non-specific protein cleavages or random instrumental errors. Efforts directed to minimize sources of error may improve the overall sensitivity by increasing the proportion of product ion mass spectra that match database peptides.

Product Ion Mass Spectral Characterization of Bacteria. The 1D-RP-LC-MS/MS analyses of tryptic peptides from bacterial cells combined with database search results of known distributions and validity allow the use of different software filtering criteria to achieve desired specificities, sensitivities, and error rates. Histograms displaying numbers of accepted peptides with known precision of correct assignments are shown in Figures 3 and 4. Precision (Pr) is defined as the fraction of correct peptide matches among unique (U) peptide assignments passing the filter [Pr = True Positive/(True Positive+False Positive)]. The horizontal dashed lines across these graphs indicate the expected number of incorrectly assigned peptides used for construction of each histogram [(1 - Pr) x U].

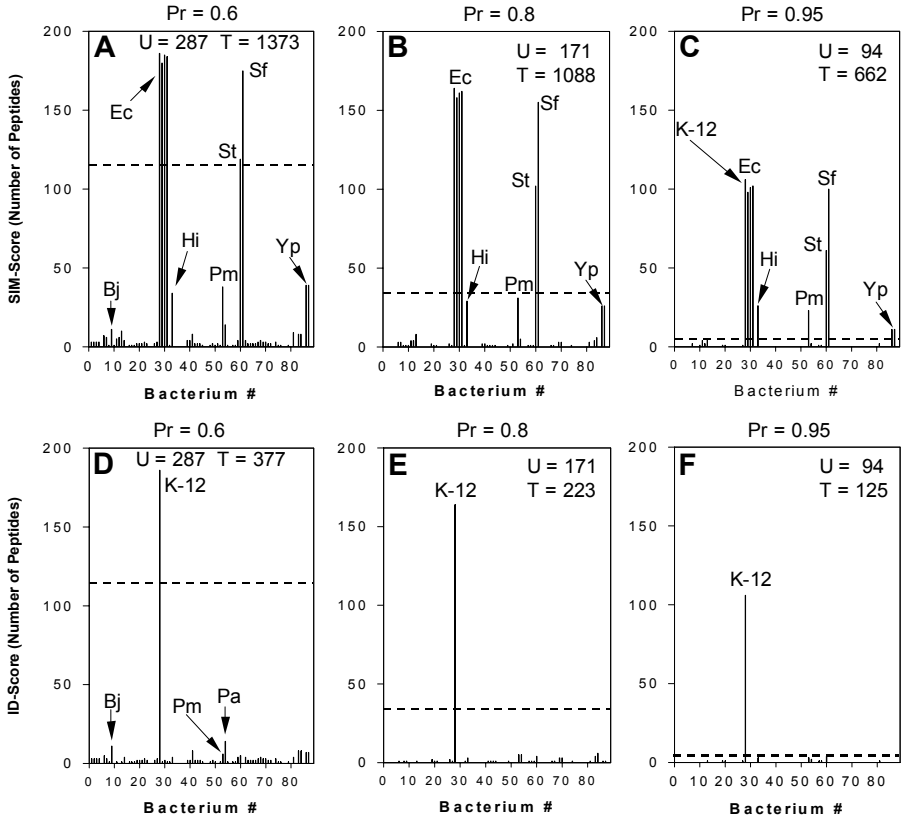

Figure 3. The number of accepted peptides from the analysis of an *E. coli* K-12 protein digest matched to translatomes of each organism in the database [Bacterium #, where consecutive numbers refer to the order of bacterial genomes Three sets of peptide assignments passing a filter (see text) at the predetermined precision levels are shown in frames A–C (plots reflect genomic similarities) while the corresponding identification plots (D–F) were obtained by filtering redundant peptides found in each set. Abbreviations: SIM-Score, similarity score reflecting the number of matching peptides assigned to each organism; ID-Score, number of non-redundant peptides assigned to each organism; Pr, precision; T, total number of accepted peptides; U, total number of unique peptides accepted; horizontal dashed line, total expected number of incorrect assignments to all organisms at the indicated precision level [(1-Pr) × U]; Ec, *Escherichia coli* strains, Bj, *Bradyrhizobium japonicum*; Hi, *Haemophilus influenzae*; Pa, *Pseudomonas aeruginosa* Pm, *Pasteurella multocida*; Sf, *Shigella flexneri*; St, *Salmonella typhimurium*; Yp, *Yersinia pestis*.

With a 0.60 precision of correct identification (Figures 3A, D), the analysis of an *E. coli* lysate resulted in the assignment of tryptic peptide product ion mass spectra to 287 unique amino acid sequences (U) in the database. Among them 154 peptides were matched to *E. coli* K-12 (53.7%) in comparison to the expected 60% (0.60) while at the 95% confidence level (Figures 3C, F) 86 of 94 peptides were matched to this strain. However, the total number of all assigned peptides (T) shown in Figures 3A–C is 5 to 7 times higher than the number of unique peptides accepted (U) due to the presence of the same amino

acid sequences (degenerate sequences) in different translatomes. Degenerate peptide sequences originate from the homologous proteins derived from related bacterial species and the data obtained reflect genomic similarities among them. To reveal a probable bacterial source of these peptides in the analyzed sample, a simple deconvolution routine can be executed using transformations to retain only unique set of peptides displayed. This algorithm is based on the reasonable assumption that organisms actually present in the analyzed sample can be associated with the highest number of matching peptides. Deconvolution was performed in an iterative way by selecting the highest scoring bacterium and filtering out peptides assigned to this organism from bins associated with all remaining strains. In the next step peptides from the second highest scoring organism can be selected in the newly assembled peptide "spectrum" and so on. In other words this procedure removes identical sequences mainly associated with orthologous proteins, that is, proteins coded by genes in separate species that are derived from the same ancestral genes or are products of the horizontal gene transfer between different strains. The outcome of these transformations gives new peptide histograms shown in Figures 3D–E.

Although only a one step deconvolution is presented in Figure 3, it is clear that the total number of distributed peptides (T) is still higher than unique ones (U). This discrepancy mainly originates from the presence of paralogous genes, i.e., related genes produced from a gene duplication event within a single genome. Moreover, the analysis of peptides from *E. coli* represents a difficult case, because the database contains four different *E. coli* strains and *S. flexneri* should be considered as another *E. coli* strain on the basis of data presented in Figure 3. This conclusion is in agreement with results based on the sequencing of eight *S. flexneri* housekeeping genes (7,160 bp) and whole genome comparisons.[12]

In the case of *B. subtilis* (data not shown), at the 95% expected confidence level of correct peptide assignments, 54 of 56 unique peptides (96.4%,) were matched to *B. subtilis* while none of the remaining virtual translatomes was associated with more than two matching sequences.

Mixture Analysis. The high resolving power of this analytical method allows the analysis of mixtures of microorganisms as documented in Figure 4. A raw output of peptide assignments generated during the analysis of a sample composed of a mixture of *E. coli* K-12 and *B. subtilis* strains is shown in Figure 4A as a histogram of 1470 peptides (800 unique). Although assignments to *E. coli* and *B. subtilis* strains predominate, a substantial number of matches is associated with *S. flexneri*, *S. typhimurium* and

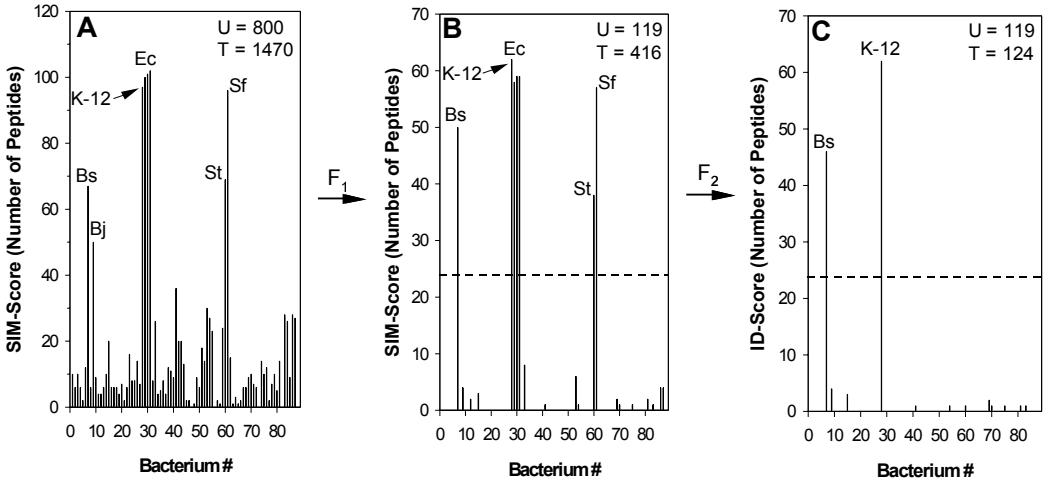

Figure 4. The number of accepted tryptic peptides from the analysis of a bacterial mixture composed of *E. coli* K-12 and *B. subtilis* (Bs) cells (2:1, w:w) whole cell digest that matched to translatomes of each organism in the database (Bacterium #): (A) raw output data from SEQUEST; (B) assignments of peptides passing the filter with the 0.8 precision level of correct identification (similarity plot); and (C) the identification plot obtained by filtering redundant peptide assignments to bacterial translatomes (for explanation, see text). For remaining abbreviations, see legend to **Figure 3**.

B. japonicum strains The substantial number of matches associated with *B. japonicum* and the lowest number of matches assigned to K-12 among *E. coli* strains reflects a substantial contribution of randomly assigned peptides in this histogram. The *B. japonicum* genome is the largest in the database (9.1 Mbp) while the K-12 genome (4.6 Mbp) is 15-20% smaller than the three other *E. coli* strains in the database. Therefore, after filtering of the low scoring assignments (Figure 4B), the number of matches associated with *B. japonicum* is diminished from 52 to 4 while K-12 clearly predominates among *E. coli* strains.

The deconvolution of the Figure 4B histogram produces the histogram shown in Figure 4C, where. the number of peptides assigned to K-12 remains the same (62) while the number of *B. subtilis* matches was diminished from 50 to 46 and reflects the presence of only four identical sequences shared by these organisms. From the total number of unique peptides (119), the peptides matching *B. subtilis* (46) and *E. coli* K-12 (57) represent 86% (the expected precision is 0.8) and only 16 assignments are distributed among the remaining organisms. However, due to the presence of paralogs, the actual number of peptides assigned to *E. coli* is 62.

CONCLUSIONS

The 45-60 minute 1D HPLC-MS/MS analysis of tryptic digests derived from pure cultures of *E. coli*, *B. thuringiensis, B. subtilis* and a bacterial mixture combined with a new data processing algorithm, which includes SEQUEST, allows for relatively fast and highly reliable identification of bacteria represented in a database in an automated fashion. The present approach is based on (a) HPLC-MS/MS analysis of microbial tryptic peptides combined with (b) searching a database composed of virtual bacterial translatomes (proteomes) and (c) an in-house developed scoring system. These procedures allow for a high throughput analysis of product ion mass spectral database search results, identification of correct peptide assignments at a chosen probability level, and high confidence level identification of pure cultures as well as mixtures of microorganisms.

Although only bacteria represented in the database can be identified, sequenced bacterial genomes include all priority pathogenic bacteria for biodefense purposes as well as their protein toxins. In addition, there is no conceptual limitation in the extension of this approach to the analysis of hundreds of viruses with sequenced genomes. Future work will include a broader range of organisms and expected environmental interferences as well as the seamless automation of the entire identification process.

ACKOWLEDGMENTS

This work was supported by the US Army Edgewood Chemical Biological Center.

REFERENCES

1. http://www.niaid.nih.gov/biodefense/bandc_priority.htm
2. Doolitle, R. F. *Nature* **2002**, *416*, 697-700.
3. Yates, J. R., 3rd.; Eng, J. K.; US Patent 6,017,693, Jan. 25, 2000.
4. Harris, W. A.; Reilly, J. P. *Anal Chem.* **2002**, *74*, 4410-4416.
5. Yao, Z-P.; Alfonso, C.; Fenselau, C. *Rapid Commun. Mass Spectrom.* **2002**, *16*, 1953-1956.
6. Link, A. J.; Eng, J.; Schieltz, D. M.; Carmack, E.; Mize, G. J.; Morris, D. R.; Garvik, B. M.; Yates, J. R., 3rd. *Nat. Biotechnol*. **1999**, *17*, 676-682.
7. Washburn, M. P.; Wolters, D.; Yates, J.R., 3rd. *Nat. Biotechnol.* **2001***, 19*, 242-247.
8. http://www.ncbi.nlm.gov/PMGifs/Genomes/micr/html.
9. Saltzber, S. L.; Delcher, A. L.; Kasif, S.; White, O. *Nucleic Acids Res.* **1998**, *26*, 544-548.
10. Eng, J. K.; McCormack, A. L.; Yates, J. R., 3rd, *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976-989.
11. Keller, A.; Nesvizhskii, A. I.; Kolker, I.; Aebersold, R. *Anal. Chem.* **2002**, *74*, 5383-5392.
12. Jin, Q.; (32 co-authors) *Nucleic Acids Res.* **2002**, *30*, 4432-4441.